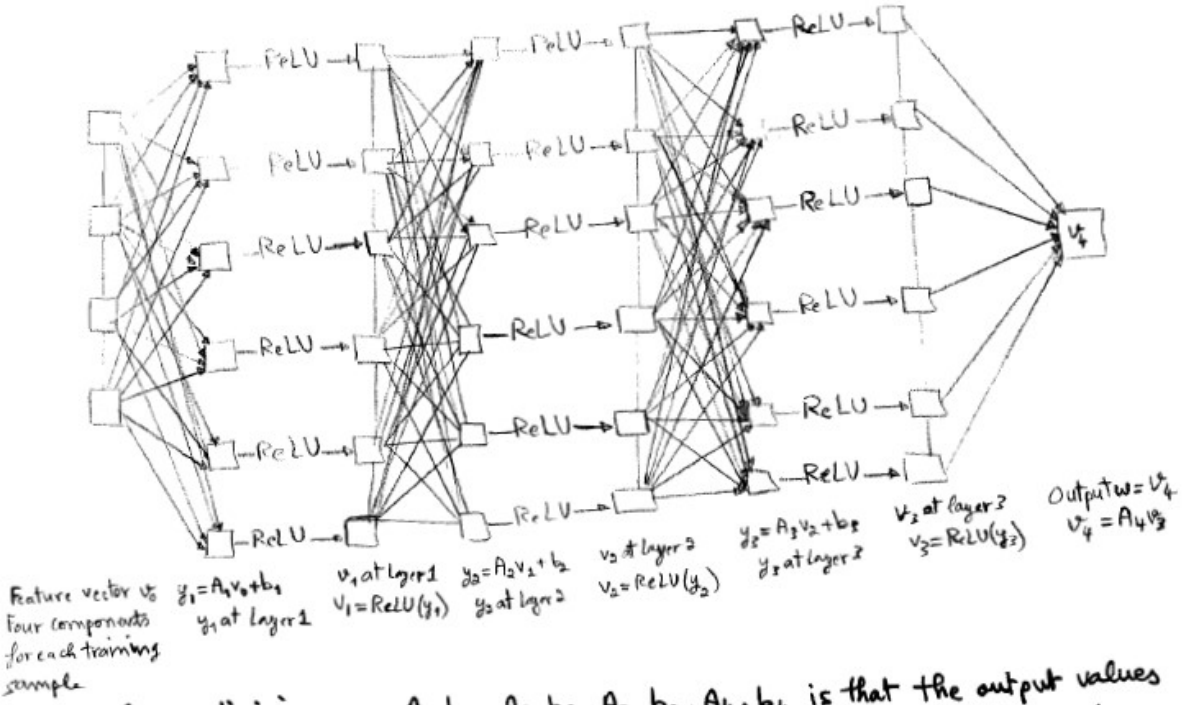
1.

problem VII.1 - 9 10/10

We have a network with $m = N_0 = 4$ inputs in each feature vector v_0 and $N = 6$ neurons on each of the 3 hidden layers

The neural network is shown below:



The goal in optimizing $x = A_1, b_1, A_2, b_2, A_3, b_3, A_4, b_4$ is that the output values $v_\ell = v_4$ at the last layer $\ell = 4$ should correctly capture the important features of the training data v_0.

A_1: 6x4 b_1: 6x1

A_2: 6x6 b_2: 6x1

A_3: 6x6 b_3: 6x1

A_4: 1x6 b_4: 1x1 (not used)

note: usually, there is no bias vector at the final step to the output (no b_4)

\# weights = 120
\# biases = 18
\# ReLU = 18

* the number of weights is the total number of elements in A_1, A_2, A_3, A_4, b_1, b_2, b_3:
$\#w = 6x4 + 2x6x6 + 1x6 + 3x6x1 = \underline{120}$
* the number of biases is the total number of elements in the bias vectors b_1, b_2, b_3:
\# biases = $3x6x1 = \underline{18}$
* the number of activation functions (ReLU):
there is one (ReLU) for each neuron on the hidden layers:
$\#(\text{ReLU}) = 6x3 = \underline{18}$

Problem VII.1 - 15

Example 4 with blue and orange spirals is much more difficult ! With one hidden layer, we explore whether the network learn this training data as N increases. We start with $N = 1$ and we go up to $N = 8$. The results are summarized as follows

Figure 2: **Example 4:** Blue and Orange Spirals, One Hidden Layer, $N = 1$

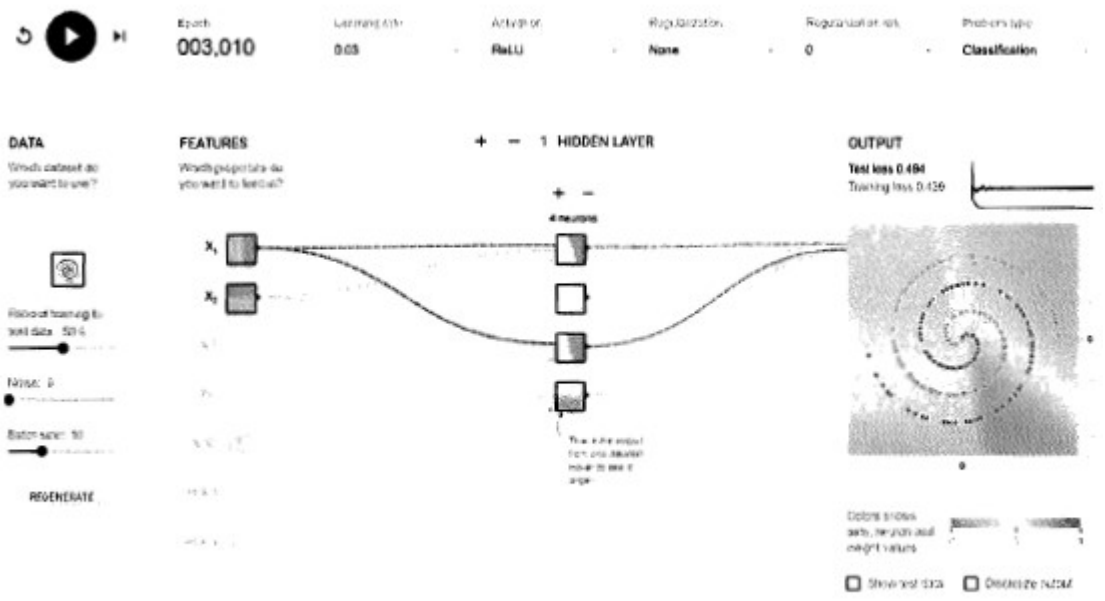


Figure 5: **Example 4:** Blue and Orange Spirals, One Hidden Layer, $N = 4$

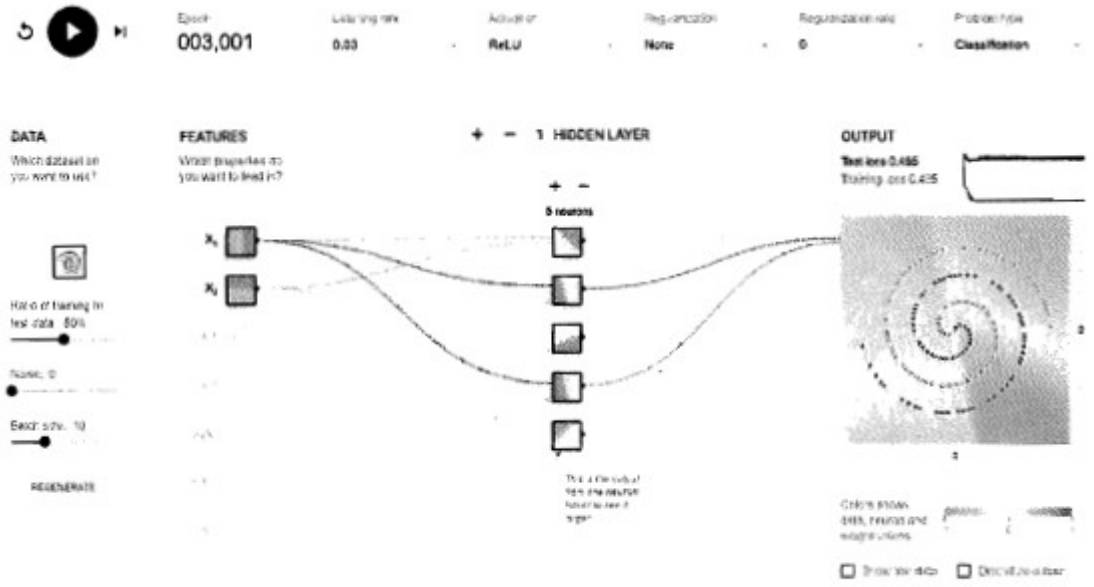


Figure 6: **Example 4:** Blue and Orange Spirals, One Hidden Layer, $N = 5$

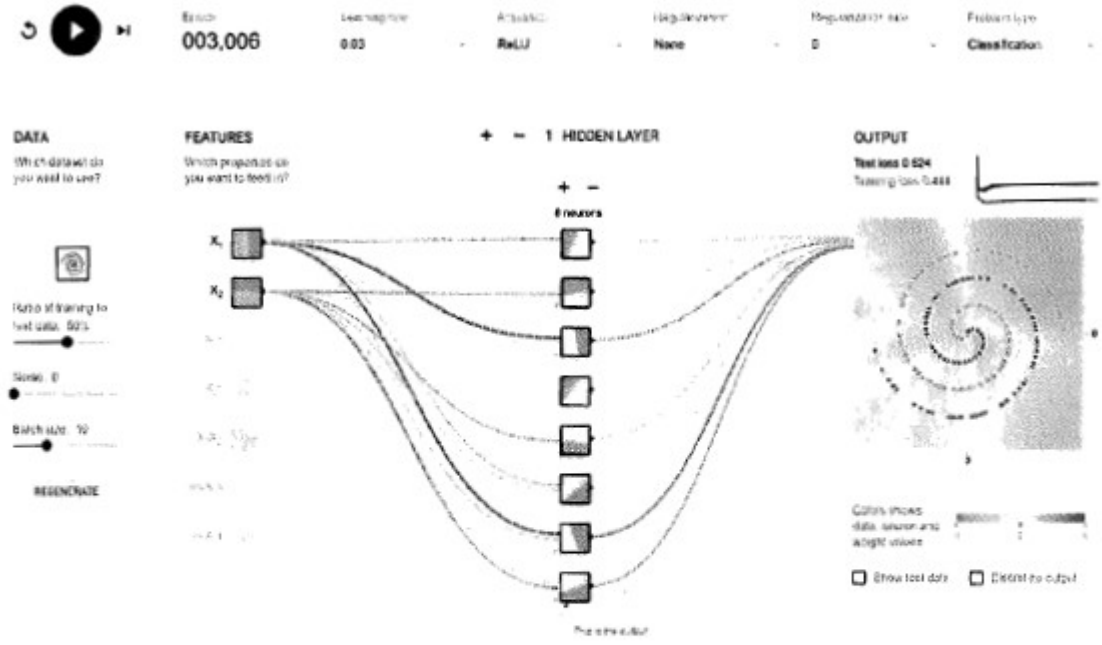


Figure 9: **Example 4:** Blue and Orange Spirals, One Hidden Layer, $N = 8$

No, the network can't learn this training data. As N increases, we observe that the network is not able to classify properly with error being almost the same. This is because the only properties (features) we are feeding in are X_1 and X_2, and we are only using one hidden layer. However, if we use two hidden layers and also feed in the two additional properties X_1^2 and X_2^2, the network is able to learn the training data as shown in Figure 13 in **Problem VII.1 - 16**.

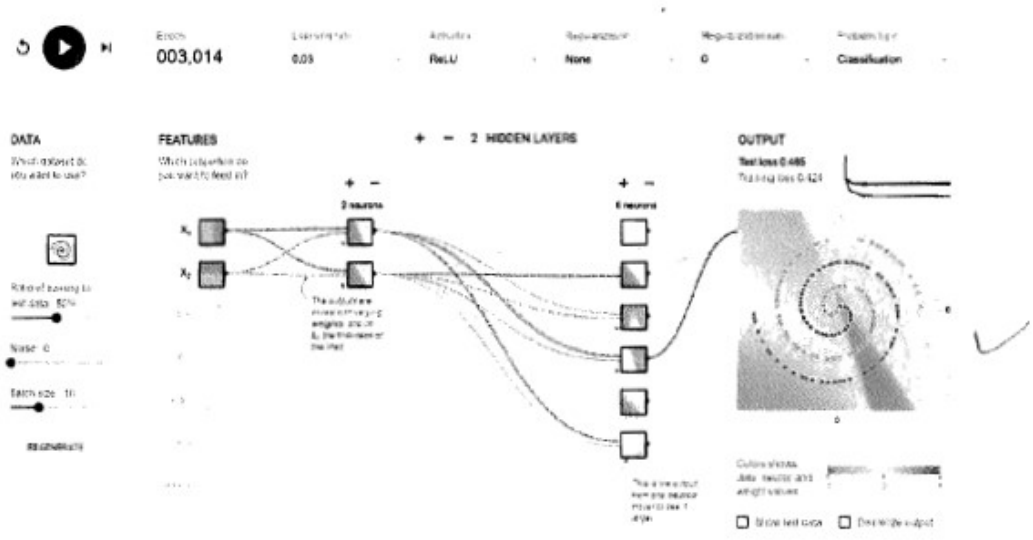


Figure 12: **Example 4:** Blue and Orange Spirals, Two Hidden Layers, $2 + 6$

As we can see in the figures above, $2 + 6$ is worse than $6 + 2$ and it is more unusual. (having higher test and training error) We note that if we use two hidden layers and also feed in the two additional properties X_1^2 and X_2^2, the network is able to learn the training data as shown in the figure below.

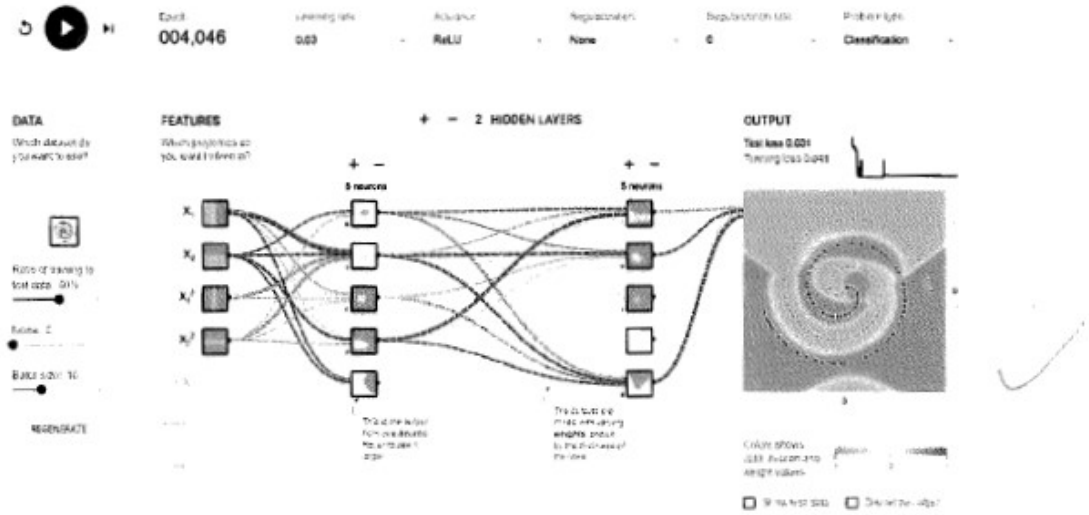


Figure 13: **Example 4:** Blue and Orange Spirals, Two Hidden Layers, $5 + 5$